Proposal for a CaBIG Proteomics Ontology

The Proteomics Workspace of caBIG Integrative Cancer Research (ICR) proposes here to develop a caBIG Proteomics Ontology in collaboration with the Vocabularies & Common Data Elements Workspace to facilitate proteomics data exchange, comparison and verification. The Proteomics Ontology should provide clear definition of collection of terms and their context for unambiguously describing proteomics experiment. This effort will establish and document the common understanding of the work flow in collecting, analyzing and comparing proteomics data, supporting the development and adoption of controlled vocabularies & common data element (CDE).

Our group has been working on developing statistical tools to analyze various proteomics data, defining standards allowing sharing, communication and comparison of experimental results, and research innovative algorithms for data mining. We have found that vast amount of proteomics experimental data exist in isolated repositories scattered around. They are often in some proprietary format and vaguely defined, therefore, very difficult to use. In addition, there is no or little help for users to evaluate the quality of data.

To exploring proteomics data from various resources, we need to define a common data model, which will enable users to describe all the relevant data pertaining to a proteomics experiment in a standardized and clear understood format. Specifically, controlled vocabularies will be required to recount parameters such as the sample source, experimental procedure, detection method, and data analysis. The data model should also be flexible enough to allow addition of user controlled vocabularies for supporting the specific needs of laboratories and diversity in experimental methodology.

There are several initiatives in developing such a community data standards including the Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) and the MIAPE Standard (Minimum Information about a proteomics experiment). However, none of them has been validated rigorously and widely adopted. The difficulty in defining and implementing proteomics standards stems from the fast advancement in proteomics technologies and the complex nature of proteomics data. For instance, multiple sites participating the HUPO Plasma Proteome Project around the world submitted Surface Enhanced Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry (SELDI-TOF MS) spectra of a set of standard plasma samples for data comparison and reproducibility. Among eight sites that turned in data, three were removed due to poor spectrum quality. For the remaining sites, the correlation in spectra across sites and across multiple samples is quite poor. It is worth noting that all submitted spectra are collected following the same requirement. To sort out all relevant variables in a proteomics experiment requires wide range of collaboration.

How to manage increasing complexity in the Controlled Vocabularies & CDEs for proteomics data? How to coordinate effort with existing initiatives? We believe the Proteomics Ontology will help to capture the workflow of proteomics experiment semantically and reconcile various existing standards. In addition, the Proteomics

Ontology will allow easy transformation to XML standards for data exchange and to UML model for application interoperation. The interpretation of biological experiment particular proteomics data increasingly requires the integration of genetic, physiological and biochemical information. We believe the ontology will become valuable by providing the knowledge base for further integration with other type of information including genomics, clinical information.